Matematikai statisztika

Programozó Matematikus szak részére

Pap Gyula
KLTE Matematikai és Informatikai Intézet
4010 Debrecen, Pf.12
papgy@math.klte.hu

1. Bevezetés

1.1. A matematikai statisztika célkitűzései

Adott egy **minta**, amely alapján következtetéseket szeretnénk levonni bizonyos valószínűségekkel, eloszlás– illetve sűrűségfüggvényekkel, paraméterekkel kapcsolatban.

Például:

Becsléselmélet: ismeretlen paraméterek közelítése

- **pontbecslés:** a mintában megállapított selejtarány alapján szeretnénk megbecsülni az egész sokaságban levő selejtarányt
- **intervallumbecslés:** konfidencia–intervallum (azaz megbízhatósági intervallum) keresése, vagyis például olyan intervallum meghatározása a mintaátlag körül, melybe a várható érték nagy valószínűséggel beleesik

Hipotézisvizsgálat: döntés bizonyos hipotézisek (feltevések) elfogadásáról vagy elvetéséről; például a mintaátlagnak a várható értéktől való eltérése belefér–e még a szabványelőírásokba?

- **illeszkedésvizsgálat:** a minta milyen eloszlásból származik, azaz milyen eloszlással közelíthető jól? (Vagyis döntés arról, hogy az adott minta illeszkedik–e az adott eloszláshoz?)
- **homogenitásvizsgálat:** döntés arról, hogy vajon két adott minta ugyanabból az eloszlásból származik–e?
- **függetlenségvizsgálat:** döntés arról, hogy vajon két adott minta független–e?

1.2. Statisztikai alapfogalmak

1.2.1. Definíció. ***(Statisztikai) minta*** *alatt valamely* $(\Omega, \mathcal{A}, \mathsf{P})$ *valószínűségi mezőn értelmezett független, azonos eloszlású valószínűségi változók* $\xi_1, \xi_2, \dots, \xi_n$ *véges sorozatát értjük. A minta* ***elemszáma*** n, *a minta* ***alapeloszlása*** *pedig a* $\xi_1, \xi_2, \dots, \xi_n$ *valószínűségi változók közös eloszlása.*

1.2.2. Definíció. ***Statisztika*** *(pontosabban statisztikai függvény) alatt a mintaelemek valamely* $T(\xi_1, \xi_2, \dots, \xi_n)$ *függvényét értjük, ahol* $T : \mathbb{R}^n \to \mathbb{R}$ *olyan függvény, hogy* $T(\xi_1, \xi_2, \dots, \xi_n)$ *valószínűségi változó.*

Megjegyezzük, hogy ehhez elegendő például az, hogy T folytonos legyen, hiszen ekkor tetszőleges $c \in \mathbb{R}$ esetén

$$\{\omega \in \Omega : T(\xi_1(\omega), \dots, \xi_n(\omega)) < c\} = \{\omega \in \Omega : (\xi_1(\omega), \dots, \xi_n(\omega)) \in T^{-1}(-\infty, c)\} \in \mathcal{A},$$

azaz esemény, ugyanis T folytonossága miatt $T^{-1}(-\infty, c) = \{x \in \mathbb{R}^n : T(x) \in (-\infty, c)\}$ nyílt halmaz, ezért előáll nyílt intervallumok megszámlálható uniójaként.

1.2.3. Definíció. ***Mintaátlag:***

$$\overline{\xi}_n := \frac{\xi_1 + \xi_2 + \dots + \xi_n}{n},$$

vagyis a mintaelemek számtani közepe.

A mintaátlag nyilván statistika, hiszen $\overline{\xi}_n = T_n(\xi_1, \xi_2, \dots, \xi_n)$, ahol

$$T_n(x_1, x_2, \dots, x_n) = \frac{x_1 + x_2 + \dots + x_n}{n}$$

folytonos függvény.

1.2.4. Definíció. ***Tapasztalati (empirikus) szórásnégyzet:***

$$s_n^2 := \frac{(\xi_1 - \overline{\xi}_n)^2 + (\xi_2 - \overline{\xi}_n)^2 + \dots + (\xi_n - \overline{\xi}_n)^2}{n} = \frac{1}{n} \sum_{i=1}^{n} \left(\xi_i - \frac{1}{n} \sum_{j=1}^{n} \xi_j \right)^2,$$

vagyis a mintaelemek átlagtól való négyzetes eltéréseinek átlaga.

A tapasztalati szórásnégyzet is statisztika, hiszen $s_n^2 = T_n(\xi_1, \xi_2, \dots, \xi_n)$, ahol

$$T_n(x_1, x_2, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right)^2$$

folytonos függvény.

1.2.5. Lemma. ***(Steiner–formula)*** *Tetszőleges* $c \in \mathbb{R}$ *esetén*

$$s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (\xi_i - c)^2 - \left(\frac{1}{n} \sum_{i=1}^{n} (\xi_i - c) \right)^2.$$

Speciálisan ($c = 0$ *választással)*

$$s_n^2 = \frac{1}{n} \sum_{i=1}^{n} \xi_i^2 - \left(\overline{\xi}_n\right)^2.$$

Bizonyítás.

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\xi_i - \overline{\xi}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left((\xi_i - c) - (\overline{\xi}_n - c)\right)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}(\xi_i - c)^2 - 2\frac{1}{n}(\overline{\xi}_n - c)\sum_{i=1}^{n}(\xi_i - c) + (\overline{\xi}_n - c)^2,$$

ahol

$$\frac{1}{n}(\overline{\xi}_n - c)\sum_{i=1}^{n}(\xi_i - c) = (\overline{\xi}_n - c)^2 = \left(\frac{1}{n}\sum_{i=1}^{n}(\xi_i - c)\right)^2$$

□

1.2.6. Definíció. *A* $\xi_1, \xi_2, \dots, \xi_n$ *mintából képezett* ***rendezett minta:***

$$\xi_{1:n}^* \leqslant \xi_{2:n}^* \leqslant \dots \leqslant \xi_{n:n}^*.$$

A rendezett minta elemei statisztikák, ugyanis például $\xi_{1:n}^* = T_n(\xi_1, \xi_2, \dots, \xi_n)$, ahol

$$T_n(x_1, x_2, \dots, x_n) = \min\{x_1, x_2, \dots, x_n\}$$

folytonos függvény. Megjegyezzük, hogy a $\xi_{1:n}^*, \xi_{2:n}^*, \dots, \xi_{n:n}^*$ valószínűségi változók nem függetlenek és nem azonos eloszlásúak.

1.2.7. Definíció. *A* $\xi_1, \xi_2, \dots, \xi_n$ *mintából képezett* ***tapasztalati medián:***

$$\begin{cases} \xi_{m+1:n}^*, & \textit{ha } \ n = 2m+1 \ \textit{ páratlan,} \\ (\xi_{m:n}^* + \xi_{m+1:n}^*)/2, & \textit{ha } \ n = 2m \ \textit{ páros.} \end{cases}$$

A $\xi_1, \xi_2, \dots, \xi_n$ *minta* ***terjedelme:***

$$\xi_{n:n}^* - \xi_{1:n}^*.$$

1.2.8. Definíció. *A* $\xi_1, \xi_2, \dots, \xi_n$ *mintából képezett* ***tapasztalati (empirikus) eloszlásfüggvény:*** $\mathcal{F}_n : \mathbb{R} \times \Omega \to \mathbb{R}$,

$$\mathcal{F}_n(x) = \mathcal{F}_n(x, \omega) := \sum_{i:\, \xi_i(\omega) < x} \frac{1}{n} = \begin{cases} 0, & \textit{ha } \ x \leqslant \xi_{1:n}^*, \\ \dfrac{k}{n}, & \textit{ha } \ \xi_{k:n}^* < x \leqslant \xi_{k+1:n}^*, \\ 1, & \textit{ha } \ x > \xi_{n:n}^*. \end{cases}$$

Vagyis minden rögzített $x \in \mathbb{R}$ esetén $\mathcal{F}_n(x, \cdot) : \Omega \to \mathbb{R}$ valószínűségi változó, ezért statisztika, mégpedig a $\{\omega \in \Omega : \xi_i(\omega) < x\} \in \mathcal{A}$ esemény relatív gyakorisága. Továbbá minden rögzített $\omega \in \Omega$ esetén $\mathcal{F}_n(\cdot, \omega) : \mathbb{R} \to \mathbb{R}$ eloszlásfüggvény, mégpedig annak a diszkrét eloszlásnak az eloszlásfüggvénye, melynek a lehetséges értékei a mintaelemek $\xi_1(\omega), \xi_2(\omega), \dots, \xi_n(\omega)$ realizációi, és ezen értékeket egyaránt $1/n$ valószínűséggel veheti fel.

1.2.9. Tétel. ***(Glivenko) (A matematikai statisztika alaptétele)*** *Legyenek* $\xi_1, \xi_2, \dots$ *független, azonos eloszlású valószínűségi változók közös* F *eloszlásfüggvénnyel. Ekkor*

$$\mathsf{P}\left(\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|\mathcal{F}_n(x) - F(x)| = 0\right) = 1.$$

Tehát a tapasztalati eloszlásfüggvény egyenletesen közelíti az (elméleti) alapeloszlást, midőn az n mintaelemszám tart végtelenbe. Megjegyezzük, hogy minden rögzített $x \in \mathbb{R}$ esetén a nagy számok erős törvénye alapján 1 valószínűséggel teljesül $\mathcal{F}_n(x) \to \mathsf{P}(\xi_1 < x) = F(x)$ ha $n \to \infty$, ezért $\mathsf{P}\left(\lim_{n\to\infty}|\mathcal{F}_n(x) - F(x)| = 0\right) = 1$.

1.2.10. Definíció. *Legyen* $r \in \mathbb{N}$. *A* $\xi_1, \xi_2, \dots, \xi_n$ *mintából képezett* ***tapasztalati r–edik momentum:***

$$m_{r,n} := \frac{1}{n}\sum_{i=1}^{n}\xi_i^r.$$

Nyilván $\overline{\xi}_n = m_{1,n}$, és a Steiner–formula alapján $s_n^2 = m_{2,n} - m_{1,n}^2$. Továbbá $m_{r,n}$ éppen az $\mathcal{F}_n$ (véletlentől függő) eloszlásfüggvényhez tartozó r–edik momentum.

Hisztogram (oszlopdiagram) úgy kapható, hogy vesszük a valós számegyenesnek egy $-\infty < x_1 < x_2 < \dots < x_k < \infty$ beosztását, és az egyes részintervallumokra olyan magasságú téglalapokat rajzolunk, ahány mintalem az illető részintervallumba esik; vagyis egy olyan lépcsősfüggvényről van szó, amelynek értéke az egyes részintervallumokon az illető részintervallumba esés gyakorisága. **Sűrűségi hisztogram** úgy kapható, hogy a relatív gyakoriságokat ábrázoljuk. Ha a mintaelemszám végtelenbe tart és a beosztássorozat finomodik, akkor a sűrűségi hisztogram jól közelíti a sűrűségfüggvényt.

2. A statisztikában használatos néhány fontos eloszlás

2.1. χ_k^2–eloszlás

2.1.1. Definíció. *Ha* $\eta_1, \eta_2, \dots, \eta_k$ *független, standard normális eloszlású valószínűségi változók, akkor*

$$\chi_k^2 := \eta_1^2 + \eta_2^2 + \dots + \eta_k^2$$

eloszlását $\boldsymbol{\chi_k^2}$***–eloszlásnak*** *nevezzük, melynek* ***szabadsági foka*** $\boldsymbol{k}$.

2.1.2. Definíció. ***Szemifaktoriális:***

$$n!! := \begin{cases} n(n-2)(n-4)\dots 1, & \textit{ha } n \textit{ páratlan,} \\ n(n-2)(n-4)\dots 2, & \textit{ha } n \textit{ páros.} \end{cases}$$

2.1.3. Lemma. *A χ_k^2–eloszlás abszolút folytonos, és sűrűségfüggvénye*

$$f_{\chi_k^2}(x) = \begin{cases} 0, & \textit{ha} \ \ x \leqslant 0, \\ c_k x^{k/2-1} \mathrm{e}^{-x/2}, & \textit{ha} \ \ x > 0, \end{cases}$$

ahol

$$c_k = \begin{cases} \dfrac{1}{\sqrt{2\pi} \cdot (k-2)!!}, & \textit{ha} \ \ k \ \ \textit{páratlan}, \\ \dfrac{1}{2 \cdot (k-2)!!}, & \textit{ha} \ \ k \ \ \textit{páros}. \end{cases}$$

Speciálisan

$$f_{\chi_1^2}(x) = \begin{cases} 0, & \text{ha} \ \ x \leqslant 0, \\ \dfrac{1}{\sqrt{2\pi x}} \mathrm{e}^{-x/2}, & \text{ha} \ \ x > 0, \end{cases} \qquad f_{\chi_2^2}(x) = \begin{cases} 0, & \text{ha} \ \ x \leqslant 0, \\ \dfrac{1}{2} \mathrm{e}^{-x/2}, & \text{ha} \ \ x > 0. \end{cases}$$

Tehát a χ_2^2–eloszlás megegyezik az $1/2$–paraméterű exponenciális eloszlással.

Bizonyítás. Nyilván χ_1^2 eloszlásfüggvénye

$$F_{\chi_1^2}(x) = \mathsf{P}(\chi_1^2 < x) = \mathsf{P}(\eta_1^2 < x) = \begin{cases} 0, & \text{ha} \ \ x \leqslant 0, \\ \mathsf{P}(|\eta_1| < \sqrt{x}), & \text{ha} \ \ x > 0. \end{cases}$$

Ha $x > 0$, akkor

$$\mathsf{P}(|\eta_1| < \sqrt{x}) = \mathsf{P}(-\sqrt{x} < \eta_1 < \sqrt{x}) = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{x}}^{\sqrt{x}} \mathrm{e}^{-u^2/2} \, \mathrm{d}u = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{x}} \mathrm{e}^{-u^2/2} \, \mathrm{d}u = G(\sqrt{x}),$$

ahol

$$G(y) = \frac{2}{\sqrt{2\pi}} \int_0^y \mathrm{e}^{u^2/2} \, \mathrm{d}u.$$

Nyilván $G''(y) = \sqrt{2/\pi} \mathrm{e}^{-y^2/2}$, ezért $F_{\chi_1^2}$ deriválható $(0, \infty)$–en:

$$F'_{\chi_1^2}(x) = G'(\sqrt{x}) \frac{1}{2\sqrt{x}} = \sqrt{\frac{2}{\pi}} \mathrm{e}^{-x/2} \frac{1}{2\sqrt{x}},$$

ezért χ_1^2 abszolút folytonos és sűrűségfüggvénye

$$f_{\chi_1^2}(x) = \begin{cases} 0, & \text{ha} \ \ x \leqslant 0, \\ \dfrac{1}{\sqrt{2\pi x}} \mathrm{e}^{-x/2}, & \text{ha} \ \ x > 0. \end{cases}$$

Továbbá χ_2^2 eloszlásfüggvénye

$$\begin{aligned} f_{\chi_2^2}(x) &= f_{\eta_1^2 + \eta_2^2}(x) = \int_{-\infty}^{\infty} f_{\eta_1^2}(y) f_{\eta_2^2}(x - y) \, \mathrm{d}y \\ &= \int_{-\infty}^{\infty} f_{\chi_1^2}(y) f_{\chi_1^2}(x - y) \, \mathrm{d}y = \begin{cases} 0, & \text{ha} \ \ x \leqslant 0, \\ \displaystyle\int_0^x \frac{1}{2\pi y} \mathrm{e}^{-y/2} \frac{1}{2\pi (x-y)} \mathrm{e}^{-(x-y)/2} \, \mathrm{d}y, & \text{ha} \ \ x > 0. \end{cases} \end{aligned}$$

Tehát ha $x > 0$, akkor

$$f_{\chi_2^2}(x) = \frac{1}{2\pi}\mathrm{e}^{-x/2}\int_0^x \frac{\mathrm{d}y}{\sqrt{y(x-y)}} = \frac{1}{2\pi}\mathrm{e}^{-x/2}\int_0^1 \frac{\mathrm{d}z}{\sqrt{z(1-z)}} = c\,\mathrm{e}^{-x/2},$$

ahol $y = xz$ helyettesítést hajtottunk végre. A c konstans abból határozható meg, hogy

$$1 = \int_{-\infty}^{\infty} f_{\chi_2^2}(x)\,\mathrm{d}x = c\int_0^{\infty} \mathrm{e}^{-x/2}\,\mathrm{d}x = 2c.$$

Tehát végülis

$$f_{\chi_2^2}(x) = \begin{cases} 0, & \text{ha } x \leqslant 0, \\ \frac{1}{2}\mathrm{e}^{-x/2}, & \text{ha } x > 0. \end{cases}$$

Az általános eset teljes indukcióval bizonyítható. Ha feltesszük, hogy az állítás igaz k–ra, akkor

$$\begin{aligned} f_{\chi_{k+1}^2}(x) &= f_{\chi_k^2+\eta_{k+1}^2}(x) = \int_{-\infty}^{\infty} f_{\chi_k^2}(y) f_{\eta_{k+1}^2}(x-y)\,\mathrm{d}y = \int_{-\infty}^{\infty} f_{\chi_k^2}(y) f_{\chi_1^2}(x-y)\,\mathrm{d}y \\ &= \begin{cases} 0, & \text{ha } x \leqslant 0, \\ \int_0^x c_k y^{k/2-1}\mathrm{e}^{-y/2} c_1 (x-y)^{-1/2}\mathrm{e}^{-(x-y)/2}\,\mathrm{d}y, & \text{ha } x > 0, \end{cases} \end{aligned}$$

így ha $x > 0$, akkor

$$\begin{aligned} f_{\chi_{k+1}^2}(x) &= c_1 c_k \mathrm{e}^{-x/2}\int_0^x y^{k/2-1}(x-y)^{-1/2}\,\mathrm{d}y \\ &= c_1 c_k \mathrm{e}^{-x/2}\int_0^1 (xz)^{k/2-1}[x(1-z)]^{-1/2} x\,\mathrm{d}z = \widetilde{c}_{k+1} x^{(k+1)/2-1}\mathrm{e}^{-x/2}, \end{aligned}$$

ahol $\widetilde{c}_{k+1}$ abból határozható meg, hogy

$$1 = \int_{-\infty}^{\infty} f_{\chi_{k+1}^2}(x)\,\mathrm{d}x = \widetilde{c}_{k+1}\int_0^{\infty} x^{(n+1)/2-1}\mathrm{e}^{-x/2}\,\mathrm{d}x.$$

Mivel

$$\begin{aligned} I_{k+1} &:= \int_0^{\infty} x^{(k+1)/2-1}\mathrm{e}^{-x/2}\,\mathrm{d}x \\ &= \left[x^{(k+1)/2-1}(-2)\mathrm{e}^{-x/2}\,\mathrm{d}x\right]_{x=0}^{x=\infty} + (k-1)\int_0^{\infty} x^{(k-1)/2-2}\mathrm{e}^{-x/2}\,\mathrm{d}x = (k-1)I_{k-1}, \end{aligned}$$

így

$$\widetilde{c}_{k+1} = \frac{1}{I_{k+1}} = \frac{1}{(k-1)I_{k-1}} = \frac{\widetilde{c}_{k-1}}{k-1},$$

amiből $c_1 = 1/\sqrt{2\pi}$ és $c_2 = 1/2$ figyelembevételével kapjuk az állítást. $\square$

2.2. t_k–eloszlás (Student–eloszlás)

2.2.1. Definíció. *Ha* η *és* χ^2_k *független valószínűségi változók standard normális, illetve* χ^2_k*–eloszlással, akkor*

$$\mathsf{t}_k := \frac{\eta}{\sqrt{\chi^2_k/k}}$$

eloszlását $\mathbf{t}_k$***–eloszlásnak (Student–eloszlásnak)*** *nevezzük, melynek* ***szabadsági foka k****.*

A t_k–eloszlás sűrűségfüggvényének meghatározásához szükségünk van a következő lemmára.

2.2.2. Lemma. *Legyenek* ξ *és* ζ *abszolút folytonos, független valószínűségi változók* f_ξ*, illetve* f_ζ *sűrűségfüggvénnyel. Ekkor* ξ/ζ *abszolút folytonos, és sűrűségfüggvénye*

$$f_{\xi/\zeta}(x) = \int_{-\infty}^{\infty} |v| f_\xi(xv) f_\zeta(v)\,\mathrm{d}v.$$

Bizonyítás. Nyilván ξ/ζ eloszlásfüggvénye

$$\begin{aligned}
F_{\xi/\zeta}(x) = \mathsf{P}(\xi/\zeta < x) &= \iint\limits_{u/v<x} f_{\xi,\zeta}(u,v)\,\mathrm{d}u\,\mathrm{d}v \\
&= \int_{-\infty}^{0} \left(\int_{xv}^{\infty} f_\xi(u) f_\zeta(v)\,\mathrm{d}u \right) \mathrm{d}v + \int_{0}^{\infty} \left(\int_{-\infty}^{xv} f_\xi(u) f_\zeta(v)\,\mathrm{d}u \right) \mathrm{d}v,
\end{aligned}$$

ezért

$$f_{\xi/\zeta}(x) = F'_{\xi/\zeta}(x) = \int_{-\infty}^{0} (-v) f_\xi(xv) f_\zeta(v)\,\mathrm{d}v + \int_{0}^{\infty} v f_\xi(xv) f_\zeta(v)\,\mathrm{d}v.$$

□

2.2.3. Lemma. *A* t_k*–eloszlás abszolút folytonos, és sűrűségfüggvénye*

$$f_{\mathsf{t}_k}(x) = \frac{\widetilde{c}_k}{\left(\dfrac{x^2}{k} + 1 \right)^{(k+1)/2}},$$

ahol

$$\widetilde{c}_k = \begin{cases} \dfrac{(k-1)!!}{\pi\sqrt{k}\cdot(k-2)!!}, & \text{ha } k \text{ páratlan,} \\[2ex] \dfrac{(k-1)!!}{2\sqrt{k}\cdot(k-2)!!}, & \text{ha } k \text{ páros.} \end{cases}$$

Speciálisan

$$f_{\mathsf{t}_1}(x) = \frac{1}{\pi(x^2+1)}, \qquad f_{\mathsf{t}_2}(x) = \frac{1}{(x^2+2)^{3/2}}.$$

Tehát a t_1–eloszlás megegyezik a Cauchy–eloszlással.

Bizonyítás. Először meghatározzuk $\sqrt{\chi_k^2/k}$ eloszlásfüggvényét:

$$F_{\sqrt{\chi_k^2/k}}(x) = \mathsf{P}(\sqrt{\chi_k^2/k} < x) = \begin{cases} 0, & \text{ha } x \leqslant 0, \\ \mathsf{P}(\chi_k^2 < kx^2), & \text{ha } x > 0. \end{cases}$$

Tehát ha $x > 0$, akkor $F_{\sqrt{\chi_k^2/k}}(x) = F_{\chi_k^2}(kx^2)$, ami deriválható: $F'_{\sqrt{\chi_k^2/k}}(x) = 2kxf_{\chi_k^2}(kx^2)$, ezért $\sqrt{\chi_k^2/k}$ abszolút folytonos, és sűrűségfüggvénye

$$f_{\sqrt{\chi_k^2/k}}(x) = \begin{cases} 0, & \text{ha } x \leqslant 0, \\ 2kxc_k(kx^2)^{k/2-1}\mathrm{e}^{-kx^2/2}, & \text{ha } x > 0, \end{cases}$$

így ha $x > 0$, akkor $f_{\sqrt{\chi_k^2/k}}(x) = 2k^{k/2}c_kx^{k-1}\mathrm{e}^{-kx^2/2}$. Ebből a 2.2.2. Lemma felhasználásával

$$\begin{aligned} f_{\mathsf{t}_k}(x) &= \int_{-\infty}^{\infty} |v| f_\eta(xv) f_{\sqrt{\chi_k^2/k}}(v)\,\mathrm{d}v = \int_0^\infty v\frac{1}{\sqrt{2\pi}}e^{-x^2v^2/2}2k^{k/2}c_kv^{k-1}\mathrm{e}^{-kv^2/2}\,\mathrm{d}v \\ &= \frac{2k^{k/2}c_k}{\sqrt{2\pi}}\int_0^\infty v^k\mathrm{e}^{-(x^2+k)v^2/2}\,\mathrm{d}v = \frac{2k^{k/2}c_k}{\sqrt{2\pi}}\int_0^\infty \frac{y^k}{(x^2+k)^{k/2}}\mathrm{e}^{-y^2/2}\frac{\mathrm{d}y}{(x^2+k)^{1/2}} \\ &= \frac{\widetilde{c}_k}{(x^2+k)^{(k+1)/2}}, \end{aligned}$$

ahol

$$\widetilde{c}_k = \frac{2k^{k/2}c_k}{\sqrt{2\pi}}\int_0^\infty y^k\mathrm{e}^{-y^2/2}\,\mathrm{d}y.$$

Ha $k = 1$, akkor $c_1 = 1/\sqrt{2\pi}$ és $\int_0^\infty y\mathrm{e}^{-y^2/2}\,\mathrm{d}y = \Big[-\mathrm{e}^{-y^2/2}\Big]_{y=0}^{y=\infty} = 1$, így $\widetilde{c}_1 = 1/\pi$, ezért $f_{\mathsf{t}_1}(x) = 1/[\pi(x^2+1)]$. Ha $k = 2$, akkor $c_2 = 1/2$ és

$$\int_0^\infty y^2\mathrm{e}^{-y^2/2}\,\mathrm{d}y = \frac{1}{2}\sqrt{2\pi}\int_{-\infty}^\infty \frac{y^2}{\sqrt{2\pi}}\mathrm{e}^{-y^2/2}\,\mathrm{d}y = \frac{\sqrt{2\pi}}{2},$$

így $\widetilde{c}_2 = 1$, ezért $f_{\mathsf{t}_2}(x) = 1/(x^2+2)^{3/2}$. Továbbá $k \geqslant 2$ esetén parciális integrálással ($y^{n-1}\cdot y\mathrm{e}^{-y^2/2}$ felbontással $\big(-\mathrm{e}^{-y^2/2}\big)' = y\mathrm{e}^{-y^2/2}$ alapján)

$$\int_0^\infty y^k\mathrm{e}^{-y^2/2}\,\mathrm{d}y = \Big[-y^{k-1}\mathrm{e}^{-y^2/2}\Big]_{y=0}^{y=\infty} + (k-1)\int_0^\infty y^{k-2}\mathrm{e}^{-y^2/2}\,\mathrm{d}y = (k-1)\int_0^\infty y^{k-2}\mathrm{e}^{-y^2/2}\,\mathrm{d}y,$$

ugyanis L'Hospital szabállyal bebizonyítható, hogy $\lim_{y\to\infty} y^{n-1}\mathrm{e}^{-y^2/2} = 0$. Ha $k \geqslant 3$ páratlan, akkor

$$\int_0^\infty y^k\mathrm{e}^{-y^2/2}\,\mathrm{d}y = (k-1)(k-3)\ldots 2\int_0^\infty y\mathrm{e}^{-y^2/2}\,\mathrm{d}y = (k-1)!!,$$

amiből

$$\widetilde{c}_k = \frac{2k^{k/2}}{\sqrt{2\pi}}\cdot\frac{(k-1)!!}{\sqrt{2\pi}\cdot(k-2)!!} = \frac{k^{k/2}\cdot(k-1)!!}{\pi\cdot(k-2)!!}.$$

Ha $k \geqslant 4$ páros, akkor

$$\int_0^\infty y^k \mathrm{e}^{-y^2/2}\,\mathrm{d}y = (k-1)(k-3)\dots 3\int_0^\infty y^2\mathrm{e}^{-y^2/2}\,\mathrm{d}y = (k-1)!!\sqrt{2\pi}/2,$$

amiből

$$\widetilde{c}_k = \frac{2k^{k/2}}{\sqrt{2\pi}}\cdot\frac{(k-1)!!\sqrt{2\pi}/2}{2\cdot(k-2)!!} = \frac{k^{k/2}\cdot(k-1)!!}{2\cdot(k-2)!!}.$$

□

2.3. $\mathsf{F}_{k,\ell}$–eloszlás

2.3.1. Definíció. *Ha χ_k^2 és χ_ℓ^2 független valószínűségi változók χ_k^2–, illetve χ_k^2–eloszlással, akkor*

$$\mathsf{F}_{k,\ell} := \frac{\chi_k^2/k}{\chi_\ell^2/\ell}$$

eloszlását $\mathsf{F}_{k,\ell}$***–eloszlásnak*** *nevezzük, melynek* ***szabadsági fokai*** $\boldsymbol{k}$ ***és*** $\boldsymbol{\ell}$.

2.3.2. Lemma. *Az $\mathsf{F}_{k,\ell}$–eloszlás abszolút folytonos, és sűrűségfüggvénye*

$$f_{\mathsf{F}_{k,\ell}}(x) = \begin{cases} 0, & \text{ha } x \leqslant 0,\\ \overline{c}_{k,\ell}\dfrac{\left(\dfrac{k}{\ell}x\right)^{k/2-1}}{\left(\dfrac{k}{\ell}x+1\right)^{(k+\ell)/2}}, & \text{ha } x > 0,\end{cases}$$

ahol

$$\overline{c}_{k,\ell} = \begin{cases} \dfrac{k\cdot(k+\ell-2)!!}{\pi\ell\cdot(k-2)!!(\ell-2)!!}, & \text{ha } k \text{ és } \ell \text{ páratlan},\\ \dfrac{k\cdot(k+\ell-2)!!}{2\ell\cdot(k-2)!!(\ell-2)!!}, & \text{ha } k \text{ és } \ell \text{ páros},\\ \dfrac{k\cdot(k+\ell-2)!!}{4\ell\cdot(k-2)!!(\ell-2)!!}, & \text{egyébként}.\end{cases}$$

Bizonyítás. Először meghatározzuk χ_n^2/n eloszlásfüggvényét:

$$F_{\chi_n^2/n}(x) = \mathsf{P}(\chi_n^2/n < x) = \mathsf{P}(\chi_n^2 < nx) = F_{\chi_n^2}(nx),$$

ami deriválható: $F'_{\chi_n^2/n}(x) = nf_{\chi_n^2}(nx)$, ezért χ_n^2/n abszolút folytonos, és sűrűségfüggvénye

$$f_{\chi_n^2/n}(x) = \begin{cases} 0, & \text{ha } x \leqslant 0,\\ nc_n(nx)^{n/2-1}\mathrm{e}^{-nx/2}, & \text{ha } x > 0.\end{cases}$$

Ebből a 2.2.2. Lemma felhasználásával

$$\begin{aligned}
f_{\mathsf{F}_{k,\ell}}(x) &= \int_{-\infty}^{\infty} |v| f_{\chi^2_k/k}(xv) f_{\chi^2_\ell/\ell}(v)\,\mathrm{d}v = \int_0^\infty v c_k k^{k/2}(xv)^{k/2-1}\mathrm{e}^{-kxv/2} c_\ell \ell^{\ell/2} v^{\ell/2-1}\mathrm{e}^{-\ell v/2}\,\mathrm{d}v \\
&= c_k c_\ell k^{k/2}\ell^{\ell/2} x^{k/2-1}\int_0^\infty v^{(k+\ell)/2-1}\mathrm{e}^{-(kx+\ell)v/2}\,\mathrm{d}v \\
&= c_k c_\ell k^{k/2}\ell^{\ell/2}\frac{x^{k/2-1}}{(kx+\ell)^{(k+\ell)/2}}\int_0^\infty y^{(k+\ell)/2-1}\mathrm{e}^{-y}\,\mathrm{d}y.
\end{aligned}$$

Ha $k+\ell$ páratlan, akkor parciális integrálással

$$\int_0^\infty y^{(k+\ell)/2-1}\mathrm{e}^{-y}\,\mathrm{d}y = \left(\frac{k+\ell}{2}-1\right)\left(\frac{k+\ell}{2}-2\right)\dots\frac{3}{2}\int_0^\infty \sqrt{y}\mathrm{e}^{-y}\,\mathrm{d}y,$$

ahol

$$\int_0^\infty \sqrt{y}\mathrm{e}^{-y}\,\mathrm{d}y = \int_0^\infty \frac{u}{\sqrt{2}}\mathrm{e}^{-u^2/2}u\,\mathrm{d}u = \frac{\sqrt{2\pi}}{2\sqrt{2}}\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-u^2/2}u\,\mathrm{d}u = \frac{\sqrt{\pi}}{2}.$$

Ha $k+\ell$ páros, akkor parciális integrálással

$$\int_0^\infty y^{(k+\ell)/2-1}\mathrm{e}^{-y}\,\mathrm{d}y = \left(\frac{k+\ell}{2}-1\right)\left(\frac{k+\ell}{2}-2\right)\dots 1\int_0^\infty \mathrm{e}^{-y}\,\mathrm{d}y,$$

ahol $\int_0^\infty \mathrm{e}^{-y}\,\mathrm{d}y = 1$. □

2.4. Normális eloszlású minta vizsgálata

2.4.1. Tétel. *Ha* $\xi_1, \xi_2, \dots, \xi_n$ *független,* $\mathcal{N}(m,\sigma^2)$*–eloszlású valószínűségi változók, akkor*

(i) $\overline{\xi}_n$ *eloszlása* $\mathcal{N}(m,\sigma^2/n)$*,*

(ii) $\overline{\xi}_n$ *és* s_n^2 *függetlenek,*

(iii) ns_n^2/σ^2 *eloszlása* χ^2_{n-1}*–eloszlás.*

Ennek a tételnek a bizonyításához felhasználjuk a karakterisztikus függvényeket.

2.4.2. Definíció. *A* $\xi:\Omega\to\mathbb{R}$ *valószínűségi változó* ***karakterisztikus függvénye*** $\varphi_\xi : \mathbb{R}\to\mathbb{C}$*,*

$$\varphi_\xi(t) := \mathsf{E}\mathrm{e}^{it\xi} := \mathsf{E}\cos(t\xi) + i\mathsf{E}\sin(t\xi).$$

A $\xi = (\xi_1,\dots,\xi_k) : \Omega\to\mathbb{R}^k$ *valószínűségi vektorváltozó* ***karakterisztikus függvénye*** *(másképpen: a* $\xi_1,\dots,\xi_k$ *valószínűségi változók együttes karakterisztikus függvénye)* $\varphi_\xi = \varphi_{\xi_1,\dots,\xi_k} : \mathbb{R}^k\to\mathbb{C}$*,*

$$\varphi_\xi(t) = \varphi_{\xi_1,\dots,\xi_k}(t_1,\dots,t_k) := \mathsf{E}\mathrm{e}^{i\langle t,\xi\rangle} = \mathsf{E}\exp\left\{i\sum_{j=1}^k t_j\xi_j\right\}.$$

2.4.3. Állítás. (i) *Ha a* $\xi = (\xi_1, \dots, \xi_k) : \Omega \to \mathbb{R}^k$ *és az* $\eta = (\eta_1, \dots, \eta_k) : \Omega \to \mathbb{R}^k$ *valószínűségi vektorváltozók függetlenek, akkor*

$$\varphi_{\xi+\eta}(t) = \varphi_\xi(t)\varphi_\eta(t), \qquad t \in \mathbb{R}^k.$$

(ii) *Ha a* $\xi = (\xi_1, \dots, \xi_k) : \Omega \to \mathbb{R}^k$ *és az* $\eta = (\eta_1, \dots, \eta_k) : \Omega \to \mathbb{R}^k$ *valószínűségi vektorváltozók karakterisztikus függvényei megegyeznek, azaz* $\varphi_\xi(t) = \varphi_\eta(t)$ *bármely* $t \in \mathbb{R}^k$ *esetén, akkor megegyeznek az eloszlásaik, azaz* $F_\xi(x) = F_\eta(x)$ *bármely* $x \in \mathbb{R}^k$ *esetén.*

(iii) *A* $\xi = (\xi_1, \dots, \xi_k) : \Omega \to \mathbb{R}^k$ *és az* $\eta = (\eta_1, \dots, \eta_\ell) : \Omega \to \mathbb{R}^\ell$ *valószínűségi vektorváltozók akkor és csak akkor függetlenek, ha*

$$\varphi_{\xi,\eta}(u, v) = \varphi_\xi(u)\varphi_\eta(v), \qquad u \in \mathbb{R}^k, \quad v \in \mathbb{R}^\ell.$$

(iv) *Ha* $\xi = (\xi_1, \dots, \xi_k) : \Omega \to \mathbb{R}^k$ *valószínűségi vektorváltozó,* A *egy* $\ell \times k$*–mátrix, és* $b \in \mathbb{R}^\ell$*, akkor*

$$\varphi_{A\xi+b}(t) = \mathrm{e}^{i\langle t,b\rangle}\varphi_\xi(A^\top t), \qquad t \in \mathbb{R}^\ell.$$

Speciálisan: ha $\xi : \Omega \to \mathbb{R}$ *valószínűségi változó és* $a, b \in \mathbb{R}$*, akkor*

$$\varphi_{a\xi+b}(t) = \mathrm{e}^{itb}\varphi_\xi(at), \qquad t \in \mathbb{R}.$$

Bizonyítás. (i). A függetlenség alapján

$$\varphi_{\xi+\eta}(t) = \mathsf{E}\mathrm{e}^{i\langle t,\xi+\eta\rangle} = \mathsf{E}\left(\mathrm{e}^{i\langle t,\xi\rangle}\mathrm{e}^{i\langle t,\eta\rangle}\right) = \mathsf{E}\mathrm{e}^{i\langle t,\xi\rangle}\mathsf{E}\mathrm{e}^{i\langle t,\eta\rangle} = \varphi_\xi(t)\varphi_\eta(t).$$

(ii). Nehéz. (Azt lehet felhasználni, hogy zárt intervallumon értelmezett folytonos függvények egyenletesen közelíthetőek trigonometrikus polinomokkal.)

(iii). Ha ξ és η függetlenek, akkor

$$\varphi_{\xi,\eta}(u, v) = \mathsf{E}\mathrm{e}^{i(\langle u,\xi\rangle+\langle v,\eta\rangle)} = \mathsf{E}\left(\mathrm{e}^{i\langle u,\xi\rangle}\mathrm{e}^{i\langle v,\eta\rangle}\right) = \mathsf{E}\mathrm{e}^{i\langle u,\xi\rangle}\mathsf{E}\mathrm{e}^{i\langle v,\eta\rangle} = \varphi_\xi(u)\varphi_\eta(v).$$

Ha pedig teljesül

$$\varphi_{\xi,\eta}(u, v) = \varphi_\xi(u)\varphi_\eta(v), \qquad u \in \mathbb{R}^k, \quad v \in \mathbb{R}^\ell,$$

akkor tekintsünk olyan $\widetilde{\xi}$ és $\widetilde{\eta}$ független valószínűségi vektorváltozókat, hogy ξ és $\widetilde{\xi}$, valamint η és $\widetilde{\eta}$ eloszlásai megegyeznek. Ekkor nyilván $\varphi_{\widetilde{\xi}} = \varphi_\xi$ és $\varphi_{\widetilde{\eta}} = \varphi_\eta$, valamint

$$\varphi_{\widetilde{\xi},\widetilde{\eta}}(u, v) = \varphi_{\widetilde{\xi}}(u)\varphi_{\widetilde{\eta}}(v) \qquad u \in \mathbb{R}^k, \quad v \in \mathbb{R}^\ell,$$

amiből következik, hogy

$$\varphi_{\xi,\eta}(u, v) = \varphi_{\widetilde{\xi},\widetilde{\eta}}(u, v) \qquad u \in \mathbb{R}^k, \quad v \in \mathbb{R}^\ell,$$

ezért a $(\xi, \eta) : \Omega \to \mathbb{R}^{k+\ell}$ és $(\widetilde{\xi}, \widetilde{\eta}) : \Omega \to \mathbb{R}^{k+\ell}$ valószínűségi vektorváltozók eloszlásai megegyeznek. Így a $\widetilde{\xi}$ és $\widetilde{\eta}$ függetlenségéből kapjuk, hogy ξ és η is függetlenek, ugyanis

$$F_{\xi,\eta}(x, y) = F_{\widetilde{\xi},\widetilde{\eta}}(x, y) = F_{\widetilde{\xi}}(x)F_{\widetilde{\eta}}(y) = F_\xi(x)F_\eta(y), \qquad x \in \mathbb{R}^k,\ y \in \mathbb{R}^\ell.$$

(iv). Nyilván
$$\varphi_{A\xi+b}(t) = \mathsf{E}\mathrm{e}^{i\langle t, A\xi+b\rangle} = \mathrm{e}^{i\langle t,b\rangle}\mathsf{E}\mathrm{e}^{i\langle A^\top t,\xi\rangle} = \mathrm{e}^{i\langle t,b\rangle}\varphi_\xi(A^\top t),$$
ugyanis $\langle t, A\xi\rangle = t^\top(A\xi) = (A^\top t)^\top\xi = \langle A^\top t, \xi\rangle$, hiszen $u, v \in \mathbb{R}^k$ esetén $\langle u, v\rangle = u^\top v$. $\square$

Példák:

- Ha η standard normális eloszlású, akkor $\varphi_\eta(t) = \mathrm{e}^{-t^2/2}$, ugyanis
$$\varphi_\eta(t) = \mathsf{E}\mathrm{e}^{it\eta} = \int_{-\infty}^{\infty} \mathrm{e}^{itx}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{e}^{-t^2/2-(x-it)^2/2}\,\mathrm{d}x,$$
és belátható, hogy $\int_{-\infty}^{\infty}\mathrm{e}^{-(x-it)^2/2}\,\mathrm{d}x = \int_{-\infty}^{\infty}\mathrm{e}^{-y^2/2}\,\mathrm{d}y = \sqrt{2\pi}$.
Ha pedig ξ normális eloszlású (m, σ^2) paraméterekkel, akkor $\eta := (\xi - m)/\sigma$ standard normális eloszlású, tehát $\xi = \sigma\eta + m$ alapján $\varphi_\xi(t) = \mathrm{e}^{itm}\mathrm{e}^{-(\sigma t)^2/2} = \mathrm{e}^{itm-\sigma^2t^2/2}$.
- Ha ξ exponenciális eloszlású λ paraméterrel, akkor $\varphi_\xi(t) = 1/(1 - it/\lambda)$, hiszen
$$\varphi_\xi(t) = \int_0^\infty \mathrm{e}^{itx}\lambda\mathrm{e}^{-\lambda x}\,\mathrm{d}x = \lambda\int_0^\infty \mathrm{e}^{(it-\lambda)x}\,\mathrm{d}x = \frac{\lambda}{(it-\lambda)}\left[\mathrm{e}^{(it-\lambda)x}\right]_{x=0}^{x=\infty} = -\frac{\lambda}{it-\lambda}.$$
- Mivel a χ_2^2–eloszlás megegyezik az $1/2$–paraméterű exponenciális eloszlással, így annak a karakterisztikus függvénye $\varphi_{\chi_2^2}(t) = 1/(1-2it)$. A χ_2^n–eloszlás definíciója alapján
$$\varphi_{\chi_2^n}(t) = \varphi_{\sum_{k=1}^n \eta_k^2}(t) = \prod_{k=1}^n \varphi_{\eta_k^2}(t) = \left(\varphi_{\eta_1^2}(t)\right)^n,$$
ezért $1/(1-2it) = \varphi_{\chi_2^2}(t) = \left(\varphi_{\eta_1^2}(t)\right)^2$ alapján $\varphi_{\eta_1^2}(t) = 1/\sqrt{1-2it}$, így végülis $\varphi_{\chi_n^2}(t) = (1-2it)^{-n/2}$.

2.4.4. Definíció. *Az $\eta = (\eta_1, \ldots, \eta_k) : \Omega \to \mathbb{R}^k$ valószínűségi vektorváltozót **k–dimenziós standard normális eloszlásúnak** nevezzük, ha $\eta_1, \ldots, \eta_k$ független, standard normális eloszlásúak.*

2.4.5. Definíció. *A $\xi = (\xi_1, \ldots, \xi_k) : \Omega \to \mathbb{R}^k$ valószínűségi vektorváltozót **k–dimenziós normális eloszlásúnak** nevezük, ha létezik olyan A $k \times k$–mátrix és $b \in \mathbb{R}^k$ vektor úgy, hogy ξ és $A\eta + b$ eloszlása megegyezik, ahol $\eta = (\eta_1, \ldots, \eta_k) : \Omega \to \mathbb{R}^k$ k–dimenziós standard normális eloszlású.*

2.4.6. Tétel. *A $\xi = (\xi_1, \ldots, \xi_k) : \Omega \to \mathbb{R}^k$ valószínűségi vektorváltozó akkor és csak akkor normális eloszlású, ha karakterisztikus függvénye*
$$\varphi_\xi(t) = \mathrm{e}^{i\langle m,t\rangle - \langle Dt,t\rangle/2}, \qquad t \in \mathbb{R}^k$$
alakú, ahol $m \in \mathbb{R}^k$, D pedig egy valós, szimmetrikus, pozitív szemidefinit $k \times k$–mátrix (azaz tetszőleges $t = (t_1, \ldots, t_k) \in \mathbb{R}^k$ esetén $\langle Dt, t\rangle = \sum_{j=1}^k\sum_{\ell=1}^k d_{j,\ell}t_jt_\ell \geqslant 0$). Továbbá $\mathsf{E}\xi = (\mathsf{E}\xi_1, \ldots, \mathsf{E}\xi_d) = m$ és $\operatorname{cov}(\xi) = \mathsf{E}\left[(\xi - \mathsf{E}\xi)(\xi - \mathsf{E}\xi)^\top\right] = D$.

Ha D invertálható, akkor ξ abszolút folytonos, és sűrűségfüggvénye

$$f_\xi(x) = \frac{1}{(2\pi)^{k/2}\sqrt{\det(D)}} \mathrm{e}^{-\langle D^{-1}(x-m),\, x-m\rangle/2}.$$

2.4.7. Tétel. *Ha $(\xi_1, \ldots, \xi_m, \eta_1, \ldots, \eta_n) : \Omega \to \mathbb{R}^{m+n}$ normális eloszlású, és $\operatorname{cov}(\xi_k, \eta_\ell) = 0$ teljesül minden $k = 1, \ldots, m$ és $\ell = 1, \ldots, n$ esetén, akkor $\xi := (\xi_1, \ldots, \xi_m)$ és $\eta := (\eta_1, \ldots, \eta_n)$ függetlenek.*

Bizonyítás. A $(\xi, \eta) := (\xi_1, \ldots, \xi_m, \eta_1, \ldots, \eta_n)$ várható érték vektora és kovarianciamátrixa

$$\begin{pmatrix} \mathsf{E}_\xi \\ \mathsf{E}_\eta \end{pmatrix}, \qquad \begin{pmatrix} D_\xi & 0 \\ 0 & D_\eta \end{pmatrix},$$

ahol $D_\xi := \operatorname{cov}(\xi)$, $D_\eta := \operatorname{cov}(\eta)$. Ezért a karakterisztikus függvénye

$$\begin{aligned} \varphi_{\xi,\eta}(u,v) &= \exp\left\{ i \left\langle \begin{pmatrix} \mathsf{E}\xi \\ \mathsf{E}\eta \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix} \right\rangle - \frac{1}{2} \left\langle \begin{pmatrix} \mathsf{D}_\xi & 0 \\ 0 & D_\eta \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix} \right\rangle \right\} \\ &= \exp\left\{ i(\langle \mathsf{E}\xi, u\rangle + \langle \mathsf{E}\eta, v\rangle) - \frac{1}{2}(\langle D_\xi u, u\rangle + \langle D_\eta v, v\rangle) \right\} = \varphi_\xi(u)\varphi_\eta(v) \end{aligned}$$

tetszőleges $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$ esetén. Ebből pedig következik a 2.4.3. Állítás (iii) pontja alapján ξ és η függetlensége. □

2.4.8. Tétel. *A $\xi = (\xi_1, \ldots, \xi_k) : \Omega \to \mathbb{R}^k$ valószínűségi vektorváltozó akkor és csak akkor normális eloszlású, ha tetszőleges $c_1, \ldots, c_k \in \mathbb{R}$ esetén $\sum_{j=1}^k c_j\xi_j$ (egydimenziós) normális eloszlású.*

Bizonyítás. Ha $\xi = (\xi_1, \ldots, \xi_k)$ normális eloszlású, akkor tetszőleges $c = (c_1, \ldots, c_k) \in \mathbb{R}^k$ esetén $\eta := \langle c, \xi\rangle = \sum_{j=1}^k c_j\xi_j$ karakterisztikus függvénye

$$\varphi_\eta(t) = \mathsf{E}\mathrm{e}^{it\eta} = \mathsf{E}\mathrm{e}^{it\langle c,\xi\rangle} = \mathsf{E}\mathrm{e}^{i\langle tc,\xi\rangle} = \varphi_\xi(tc) = \mathrm{e}^{i\langle m,tc\rangle - \langle D(ct),ct\rangle/2} = \mathrm{e}^{i\langle m,c\rangle t - \langle Dc,c\rangle t^2/2},$$

tehát η valóban normális eloszlású $(\langle m, c\rangle, \langle Dc, c\rangle)$ paraméterekkel.

Ha pedig tetszőleges $c_1, \ldots, c_k \in \mathbb{R}$ esetén $\eta := \langle c, \xi\rangle = c^\top\xi$ normális eloszlású, akkor nyilván a paraméterei $\mathsf{E}\eta = c^\top\mathsf{E}\xi$ és

$$\begin{aligned} \mathsf{D}^2\eta &= \mathsf{E}\left[(\eta - \mathsf{E}\eta)(\eta - \mathsf{E}\eta)^\top\right] = \mathsf{E}\left[c^\top(\xi - \mathsf{E}\xi)\left(c^\top(\xi - \mathsf{E}\xi)\right)^\top\right] \\ &= c^\top\mathsf{E}\left[(\xi - \mathsf{E}\xi)(\xi - \mathsf{E}\xi)^\top\right] c = c^\top\operatorname{cov}(\xi)\, c. \end{aligned}$$

Így $\xi = (\xi_1, \ldots, \xi_k)$ karakterisztikus függvénye

$$\varphi_\xi(t) = \mathsf{E}\mathrm{e}^{i\langle t,\xi\rangle} = \varphi_{\langle t,\xi\rangle}(1) = \mathrm{e}^{it^\top\mathsf{E}\xi - t^\top\operatorname{cov}(\xi)t/2} = \mathrm{e}^{i\langle \mathsf{E}\xi,t\rangle - \langle\operatorname{cov}(\xi)t,t\rangle/2}$$

tetszőleges $t \in \mathbb{R}^k$ esetén, tehát ξ valóban normális eloszlású. □

A 2.4.1. Tétel bizonyítása. (i). Nyilván $\xi := (\xi_1, \ldots, \xi_n)$ normális eloszlású, hiszen $\eta_k := (\xi_k - m)/\sigma$, $k = 1, \ldots, n$ független, standard normális eloszlásúak, és $\xi_k = \sigma\eta_k + m$, $k = 1, \ldots, n$ alapján a ξ vektor előállítható, mint az $\eta := (\eta_1, \ldots, \eta_n)$ standard normális eloszlású vektor lineáris transzformáltja. Továbbá $\overline{\xi}_n$ a $\xi_1, \ldots, \xi_n$ lineáris kombinációja, ezért a 2.4.8. Tétel alapján $\overline{\xi}_n$ valóban normális eloszlású. Továbbá nyilván $\mathsf{E}\overline{\xi}_n = \sum_{j=1}^n \mathsf{E}\xi_j/n = nm/n = m$ és $\mathsf{D}^2\overline{\xi}_n = \sum_{j=1}^n \mathsf{D}^2\xi_j/n^2 = n\sigma^2/n^2 = \sigma^2/n$.

(ii). Először a 2.4.8. Tétel alapján belátható, hogy a

$$(\xi_1 - \overline{\xi}_n, \xi_2 - \overline{\xi}_n, \ldots, \xi_n - \overline{\xi}_n, \overline{\xi}_n)$$

vektor normális eloszlású, hiszen a koordinátáinak lineáris kombinációi egyúttal a normális eloszlású $(\xi_1, \xi_2, \ldots, \xi_n)$ vektor koordinátáiból képezett lineáris kombinációknak is tekinthetők.

Továbbá minden $k = 1, \ldots, n$ esetén teljesül $\operatorname{cov}(\xi_k - \overline{\xi}_n, \overline{\xi}_n) = 0$, hiszen

$$\operatorname{cov}(\xi_k - \overline{\xi}_n, \overline{\xi}_n) = \operatorname{cov}(\xi_k, \overline{\xi}_n) - \operatorname{cov}(\overline{\xi}_n, \overline{\xi}_n) = \operatorname{cov}\left(\xi_k, \frac{1}{n}\sum_{i=1}^n \xi_i\right) - \operatorname{cov}\left(\frac{1}{n}\sum_{i=1}^n \xi_i, \frac{1}{n}\sum_{j=1}^n \xi_j\right)$$
$$= \frac{1}{n}\sum_{i=1}^n \operatorname{cov}(\xi_k, \xi_i) - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \operatorname{cov}(\xi_i, \xi_j) = \frac{\sigma^2}{n} - \frac{n\sigma^2}{n^2} = 0,$$

ugyanis a $\xi_1, \xi_2, \ldots, \xi_n$ változók függetlensége miatt

$$\operatorname{cov}(\xi_i, \xi_j) = \begin{cases} \sigma^2, & \text{ha } i = j, \\ 0, & \text{ha } i \neq j. \end{cases}$$

Ezért a 2.4.7. Tétel alapján $(\xi_1 - \overline{\xi}_n, \xi_2 - \overline{\xi}_n, \ldots, \xi_n - \overline{\xi}_n)$ és $\overline{\xi}_n$ függetlenek, amiből már következik

$$s_n^2 = \frac{1}{n}\sum_{k=1}^n (\xi_k - \overline{\xi}_n)^2$$

és $\overline{\xi}_n$ függetlensége.

(iii). A Steiner–formula alkalmazásával

$$\frac{n}{\sigma^2}s_n^2 = \frac{1}{\sigma^2}\sum_{k=1}^n (\xi_k - \overline{\xi}_n)^2 = \sum_{k=1}^n \left(\frac{\xi_k - m}{\sigma}\right)^2 - \frac{1}{n}\left(\sum_{k=1}^n \frac{\xi_k - m}{\sigma}\right)^2 = \sum_{k=1}^n \eta_k^2 - \eta^2,$$

ahol $\eta_k := (\xi_k - m)/\sigma$, $k = 1, \ldots, n$ független, standard normális eloszlásúak, és

$$\eta := \frac{1}{\sqrt{n}}\sum_{k=1}^n \frac{\xi_k - m}{\sigma} = \frac{\overline{\xi}_n - m}{\sigma/\sqrt{n}}$$

is standard normális eloszlású. Tehát

$$\frac{n}{\sigma^2}s_n^2 + \eta^2 = \sum_{k=1}^n \eta_k^2,$$

ahol a baloldalon álló összeadandók (ii) alapján függetlenek. Ezért a baloldal karakterisztikus függvénye egyrészt

$$\varphi_{ns_n^2/\sigma^2+\eta^2}(t) = \varphi_{ns_n^2/\sigma^2}(t)\varphi_{\eta^2}(t),$$

másrészt pedig

$$\varphi_{ns_n^2/\sigma^2+\eta^2}(t) = \varphi_{\sum_{k=1}^n \eta_k^2}(t) = \prod_{k=1}^n \varphi_{\eta_k^2}(t) = (\varphi_{\eta^2}(t))^n .$$

Mivel tetszőleges $t \in \mathbb{R}$ esetén $\varphi_{\eta^2}(t) = 1/\sqrt{1-2it} \neq 0$, ezért lehet vele osztani, így

$$\varphi_{ns_n^2/\sigma^2}(t) = (\varphi_{\eta^2}(t))^{n-1} = \prod_{k=1}^{n-1} \varphi_{\eta_k^2}(t) = \varphi_{\sum_{k=1}^{n-1} \eta_k^2}(t),$$

tehát ns_n^2/σ^2 valóban χ_{n-1}^2–eloszlású. □

3. Becsléselmélet

Feladat: a minta alapeloszlásához rendelt valamely mennyiség becslése, azaz közelítése a mintaelemek felhasználásával, vagyis statisztikákkal.

3.1. A várható érték és a szórásnégyzet becslése

Jelölje az alapeloszlás (vagyis a $\xi_1, \xi_2, \ldots, \xi_n$ mintaelemek közös eloszlásának) várható értékét m, szórásnégyzetét σ^2. Nyilván

$$\mathsf{E}\overline{\xi}_n = \frac{1}{n}(\mathsf{E}\xi_1 + \cdots + \mathsf{E}\xi_n) = m,$$

vagyis a mintaátlag várható értéke megegyezik az alapeloszlás várható értékével. Ezért azt mondjuk, hogy a mintaátlag **torzítatlan becslése** a várható értéknek.

Továbbá

$$\mathsf{D}^2\overline{\xi}_n = \frac{1}{n^2}(\mathsf{D}^2\xi_1 + \cdots + \mathsf{D}^2\xi_n) = \frac{\sigma^2}{n} \to 0 \qquad \text{ha } n \to \infty.$$

A nagy számok erős törvénye értelmében $\mathsf{P}(\lim_{n\to\infty} \overline{\xi}_n = m) = 1$, amit úgy fejezünk ki, hogy a mintaátlag **erősen konzisztens becslése** a várható értéknek.

A centrális határeloszlátétel alapján $\sqrt{n}(\overline{\xi}_n - m)/\sigma \to \mathcal{N}(0,1)$ eloszlásban ha $n \to \infty$, hiszen

$$\lim_{n\to\infty} \mathsf{P}\left(\frac{\overline{\xi}_n - m}{\sigma/\sqrt{n}} < x\right) = \lim_{n\to\infty} \mathsf{P}\left(\frac{S_n - \mathsf{E}S_n}{\sqrt{\mathsf{D}^2 S_n}} < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \mathrm{e}^{u^2/2}\,\mathrm{d}u,$$

ahol $S_n = \xi_1 + \cdots + \xi_n$, $\mathsf{E}S_n = nm$, $\mathsf{D}^2 S_n = n\sigma^2$.

A tapasztalati szórásnégyzet várható értéke

$$\mathsf{E}s_n^2 = \frac{1}{n}\sum_{k=1}^{n}\mathsf{E}(\xi_k - \overline{\xi}_n)^2 = \frac{1}{n}\sum_{k=1}^{n}\mathsf{D}^2(\xi_k - \overline{\xi}_n) = \frac{n-1}{n}\sigma^2,$$

hiszen

$$\mathsf{D}^2(\xi_k - \overline{\xi}_n) = \mathsf{D}^2\left[\left(1-\frac{1}{n}\right)\xi_k + \frac{1}{n}\sum_{\{j:j\neq k\}}\xi_j\right] = \frac{(n-1)^2}{n^2}\mathsf{D}^2\xi_k + \frac{1}{n^2}\sum_{\{j:j\neq k\}}\mathsf{D}^2\xi_j = \frac{n-1}{n}\sigma^2.$$

Ezért a tapasztalati szórásnégyzet nem torítatlan becslése a szórásnégyetnek. Viszont a **korrigált tapasztalati szórásnégyzet:**

$$s_n^{*2} := \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \overline{\xi}_n)^2 = \frac{n}{n-1}s_n^2$$

nyilván torzítatlan becslése a szórásnégyetnek: $\mathsf{E}s_n^{*2} = \sigma^2$. Be lehet látni, hogy

$$\mathsf{D}^2\left(s_n^{*2}\right) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right),$$

ahol $\mu_4 := \mathsf{E}\left[(\xi_1 - m)^4\right]$ az alapeloszlás negyedik centrális momentuma. Azt is be lehet látni, hogy a korrigált tapasztalati szórásnégyzet erősen konzisztens becslése a szórásnégyetnek: $\mathsf{P}\left(\lim_{n\to\infty} s_n^{*2} = \sigma^2\right) = 1$.

3.2. Pontbecslések

Feltesszük, hogy az alapeloszlás, melyből a minta származik, valamely paraméteresen megadott $\{F^{(\gamma)} : \gamma \in \Gamma\}$ eloszláscsaládból való, ahol $\Gamma \subset \mathbb{R}^k$. Feladat: az ismeretlen γ paraméter becslése.

A **maximum likelihood módszer** szerint azzal a paraméterértékkel közelítünk, mely esetén a legvalószínűbb az az esemény, hogy éppen az adott mintaelemeket kapjuk.

Példák:

(i) Tegyük fel, hogy az alapeloszlás $\lambda > 0$ paraméterű Poisson–eloszlás, ahol λ ismeretlen paraméter, azaz $\xi_1, \xi_2, \ldots, \xi_n$ független valószínűségi változók, és

$$\mathsf{P}_\lambda(\xi_i = k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}, \qquad k = 0, 1, \ldots,$$

és a feladat: λ becslése a $\xi_1, \xi_2, \ldots, \xi_n$ minta alapján. Jelölje $k_1, k_2, \ldots, k_n$ a megfigyelt értékeket. A λ paraméter $\widehat{\lambda}_n$ maximum likelihood becslése olyan λ szám, ahol az

$$L_n(k_1, \ldots, k_n; \lambda) := \mathsf{P}_\lambda(\xi_1 = k_1, \ldots, \xi_n = k_n)$$

likelihood függvénynek globális maximumhelye van. Nyilván

$$L_n(k_1, \ldots, k_n; \lambda) = \prod_{i=1}^{n} \mathsf{P}_\lambda(\xi_i = k_i) = \prod_{i=1}^{n} \left(\frac{\lambda^{k_i}}{k_i!} \mathrm{e}^{-\lambda} \right) = \mathrm{e}^{-n\lambda} \prod_{i=1}^{n} \frac{\lambda^{k_i}}{k_i!},$$

és a logaritmusfüggvény monotonitása miatt elegendő megkeresni a **log–likelihood függvény:**

$$\log L_n(k_1, \ldots, k_n; \lambda) = -n\lambda + \sum_{i=1}^{n} k_i \log \lambda - \sum_{i=1}^{n} \log(k_i!)$$

globális maximumhelyét. Mivel a

$$\frac{\partial \log L_n(k_1, \ldots, k_n; \lambda)}{\partial \lambda} = -n + \sum_{i=1}^{n} \frac{k_i}{\lambda} = 0$$

egyenlet egyetlen megoldása $\lambda = \sum_{i=1}^{n} k_i/n$, és

$$\frac{\partial^2 \log L_n(k_1, \ldots, k_n; \lambda)}{\partial \lambda^2} = -\sum_{i=1}^{n} \frac{k_i}{\lambda^2} < 0$$

ha a $k_1, k_2, \ldots, k_n$ megfigyelt értékek nem mindegyike 0, így ekkor a λ paraméter maximum likelihood becslése $\widehat{\lambda}_n = \overline{\xi}_n$.

(ii) Legyen az alapeloszlás $\lambda > 0$ paraméterű exponenciális eloszlás, ahol λ ismeretlen paraméter, azaz $\xi_1, \xi_2, \ldots, \xi_n$ független valószínűségi változók melyeknek a sűrűségfüggvénye

$$f_{\xi_i}^{(\lambda)}(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & \text{ha } x > 0, \\ 0, & \text{ha } x \leqslant 0, \end{cases}$$

és a feladat: λ becslése a $\xi_1, \xi_2, \ldots, \xi_n$ minta alapján. Jelölje $x_1, x_2, \ldots, x_n$ a megfigyelt értékeket. A λ paraméter $\widehat{\lambda}_n$ maximum likelihood becslése olyan x szám, ahol az

$$L_n(x_1, \ldots, x_n; \lambda) := f_{\xi_1, \ldots, \xi_n}^{(\lambda)}(x_1, \ldots, x_n)$$

likelihood függvénynek globális maximumhelye van. Nyilván

$$L_n(x_1, \ldots, x_n; \lambda) = \prod_{i=1}^{n} f_{\xi_i}^{(\lambda)}(x_i) = \begin{cases} \lambda^n \mathrm{e}^{-\lambda(x_1 + \cdots + x_n)}, & \text{ha } x_1 > 0, \ldots, x_n > 0, \\ 0, & \text{egyébként}, \end{cases}$$

és a logaritmusfüggvény monotonitása miatt elegendő megkeresni a **log–likelihood függvény:**

$$\log L_n(x_1, \ldots, x_n; \lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

globális maximumhelyét. Mivel a

$$\frac{\partial \log L_n(x_1, \ldots, x_n; \lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

egyenlet egyetlen megoldása $\lambda = n / \sum_{i=1}^{n} x_i$, és

$$\frac{\partial^2 \log L_n(k_1, \ldots, k_n; \lambda)}{\partial \lambda^2} = -\frac{n}{\lambda^2} < 0,$$

ezért a λ paraméter maximum likelihood becslése $\widehat{\lambda}_n = 1/\overline{\xi}_n$.

(iii) Legyen az alapeloszlás egyenletes a $[0, \alpha]$ intervallumon, ahol $\alpha > 0$ az ismeretlen paraméter, azaz $\xi_1, \xi_2, \ldots, \xi_n$ független valószínűségi változók melyeknek a sűrűségfüggvénye

$$f_{\xi_i}^{(\alpha)}(x) = \begin{cases} 1/\alpha, & \text{ha } x \in [0, \alpha], \\ 0, & \text{egyébként}, \end{cases}$$

és a feladat: α becslése a $\xi_1, \xi_2, \ldots, \xi_n$ minta alapján. Az α paraméter $\widehat{\alpha}_n$ maximum likelihood becslése olyan x szám, ahol az

$$L_n(x_1, \ldots, x_n; \alpha) := f_{\xi_1, \ldots, \xi_n}^{(\alpha)}(x_1, \ldots, x_n)$$

likelihood függvénynek globális maximumhelye van. Nyilván

$$L_n(x_1, \ldots, x_n; \alpha) = \prod_{i=1}^{n} f_{\xi_i}^{(\alpha)}(x_i) = \begin{cases} 1/\alpha^n, & \text{ha } x_1 \in [0, \alpha], \ldots, x_n \in [0, \alpha], \\ 0, & \text{egyébként}, \end{cases}$$

és $x_1 \in [0, \alpha], \ldots, x_n \in [0, \alpha]$ azzal ekvivalens, hogy $\alpha \geqslant \max\limits_{1 \leqslant i \leqslant n} x_i$ ezért a globális maximumhely $\alpha = \max\limits_{1 \leqslant i \leqslant n} x_i$, így az α paraméter maximum likelihood becslése $\widehat{\alpha}_n = \max\limits_{1 \leqslant i \leqslant n} \xi_i = \xi_{n:n}^*$.

3.3. Intervallumbecslések

Megint feltesszük, hogy az alapeloszlás, melyből a minta származik, valamely paraméteresen megadott $\{F^{(\gamma)} : \gamma \in \Gamma\}$ eloszláscsaládból való, ahol $\Gamma \subset \mathbb{R}^k$. Feladat: előre megadott $\varepsilon \in (0, 1)$ számhoz olyan $S(\xi_1, \ldots, \xi_n)$ és $T(\xi_1, \ldots, \xi_n)$ statisztikákat találni, hogy

$$\mathsf{P}_\gamma\big(S(\xi_1, \ldots, \xi_n) \leqslant \gamma \leqslant T(\xi_1, \ldots, \xi_n)\big) = 1 - \varepsilon$$

teljesüljön. Ekkor azt mondjuk, hogy $\big[S(\xi_1, \ldots, \xi_n), T(\xi_1, \ldots, \xi_n)\big]$ egy **$1 - \varepsilon$ szintű konfidencia–intervallum**.

3.3.1. Konfidencia–intervallum a normális eloszlás várható értékére, ha ismert a szórásnégyzet

Tegyük fel, hogy a minta $\mathcal{N}(m, \sigma^2)$–eloszlásból származik. Ebben az esetben a 2.4.1. Tétel alapján $\overline{\xi}_n \sim \mathcal{N}(m, \sigma^2/n)$, ezért

$$\frac{\overline{\xi}_n - m}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Válasszuk meg az $u_{\varepsilon/2} \in (0,\infty)$ számot úgy, hogy ha $\eta \sim \mathcal{N}(0,1)$, akkor $\mathsf{P}(\eta > u_{\varepsilon/2}) = \varepsilon/2$, azaz $\mathsf{P}(\eta < u_{\varepsilon/2}) = 1-\varepsilon/2$, vagyis $\Phi(u_{\varepsilon/2}) = 1-\varepsilon/2$, ahol

$$\Phi(x) := F_\eta(x) = \mathsf{P}(\eta < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-u^2/2}\,\mathrm{d}u.$$

Ekkor persze a szimmetria miatt $\mathsf{P}(\eta < -u_{\varepsilon/2}) = \varepsilon/2$, ezért

$$\mathsf{P}(\eta \in [-u_{\varepsilon/2}, u_{\varepsilon/2}]) = 1 - \mathsf{P}(\eta > u_{\varepsilon/2}) - \mathsf{P}(\eta < -u_{\varepsilon/2}) = 1-\varepsilon$$

így

$$\mathsf{P}_m\left(\frac{\overline{\xi}_n - m}{\sigma/\sqrt{n}} \in [-u_{\varepsilon/2}, u_{\varepsilon/2}]\right) = 1-\varepsilon,$$

vagyis

$$\mathsf{P}_m\left(m \in \left[\,\overline{\xi}_n - u_{\varepsilon/2}\sigma/\sqrt{n},\, \overline{\xi}_n + u_{\varepsilon/2}\sigma/\sqrt{n}\,\right]\right) = 1-\varepsilon.$$

Tehát $\left[\,\overline{\xi}_n - u_{\varepsilon/2}\sigma/\sqrt{n},\, \overline{\xi}_n + u_{\varepsilon/2}\sigma/\sqrt{n}\,\right]$ egy $1-\varepsilon$ szintű konfidencia–intervallum az m paraméterre.

3.3.2. Konfidencia–intervallum a normális eloszlás várható értékére, ha nem ismert a szórásnégyzet

Tegyük fel, hogy a minta $\mathcal{N}(m,\sigma^2)$–eloszlásból származik. Ebben az esetben a 2.4.1. Tétel alapján

$$\frac{\overline{\xi}_n - m}{s_n^*/\sqrt{n}} = \frac{\dfrac{\overline{\xi}_n - m}{\sigma/\sqrt{n}}}{\sqrt{s_n^{*2}/\sigma^2}} \sim \mathsf{t}_{n-1},$$

azaz t_{n-1}–eloszlású. Válasszuk most a $t_{\varepsilon/2} \in (0,\infty)$ számot úgy, hogy $\mathsf{P}(\mathsf{t}_{n-1} > t_{\varepsilon/2}) = \varepsilon/2$ teljesüljön. Ekkor persze a szimmetria miatt $\mathsf{P}(\mathsf{t}_{n-1} < -u_{\varepsilon/2}) = \varepsilon/2$, ezért

$$\mathsf{P}(\mathsf{t}_{n-1} \in [-t_{\varepsilon/2}, t_{\varepsilon/2}]) = 1 - \mathsf{P}(\mathsf{t}_{n-1} > t_{\varepsilon/2}) - \mathsf{P}(\mathsf{t}_{n-1} < -t_{\varepsilon/2}) = 1-\varepsilon,$$

így

$$\mathsf{P}_{m,\sigma^2}\left(\frac{\overline{\xi}_n - m}{s_n^*/\sqrt{n}} \in [-t_{\varepsilon/2}, t_{\varepsilon/2}]\right) = 1-\varepsilon,$$

vagyis

$$\mathsf{P}_{m,\sigma^2}\left(m \in \left[\,\overline{\xi}_n - t_{\varepsilon/2}s_n^*/\sqrt{n},\, \overline{\xi}_n + t_{\varepsilon/2}s_n^*/\sqrt{n}\,\right]\right) = 1-\varepsilon.$$

Tehát $\left[\,\overline{\xi}_n - t_{\varepsilon/2}s_n^*/\sqrt{n},\, \overline{\xi}_n + t_{\varepsilon/2}s_n^*/\sqrt{n}\,\right]$ egy $1-\varepsilon$ szintű konfidencia–intervallum az m paraméterre.

3.3.3. Konfidencia–intervallum a normális eloszlás szórásnégyzetére

Tegyük fel, hogy a minta $\mathcal{N}(m, \sigma^2)$–eloszlásból származik. Ebben az esetben a 2.4.1. Tétel alapján ns_n^2/σ^2 egy χ^2_{n-1}–eloszlású valószínűségi változó. Válasszunk most olyan $0 < c_1 < c_2$ számokat, hogy $\mathsf{P}(\chi^2_{n-1} < c_1) = \mathsf{P}(\chi^2_{n-1} > c_2) = \varepsilon/2$ teljesüljön. Ekkor

$$\mathsf{P}_{m,\sigma^2}\left(ns_n^2/\sigma^2 \in [c_1, c_2]\right) = 1 - \varepsilon,$$

vagyis

$$\mathsf{P}_{m,\sigma^2}\left(\sigma^2 \in \left[ns_n^2/c_2,\, ns_n^2/c_1\right]\right) = 1 - \varepsilon.$$

Tehát $[ns_n^2/c_2,\, ns_n^2/c_1]$ egy $1-\varepsilon$ szintű konfidencia–intervallum a σ^2 paraméterre.

4. Statisztikai hipotézisek vizsgálata

4.1. Egymintás u–próba

Tegyük fel, hogy $\xi_1, \xi_2, \dots, \xi_n$ független, $\mathcal{N}(m, \sigma^2)$–eloszlású valószínűségi változók, ahol σ^2 ismert, m ismeretlen. Legyen $m_0 \in \mathbb{R}$ egy rögzített szám. Feladat: döntsük el, hogy a minta alapján elfogadható–e az a nullhipotézis (azaz feltevés), hogy $m = m_0$, vagy inkább az $m \neq m_0$ ellenhipotézis (alternatíva) fogadható el? Tehát:

$$\begin{cases} H_0: & m = m_0 \quad \textbf{nullhipotézis,} \\ H_1: & m \neq m_0 \quad \textbf{alternatíva}. \end{cases}$$

Ha a H_0 nullhipotézis igaz, akkor a 2.4.1. Tétel alapján

$$\mathsf{u} := \frac{\overline{\xi}_n - m_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Legyen $\varepsilon \in (0,1)$ rögzített. Válasszuk meg az $u_{\varepsilon/2} \in (0, \infty)$ számot úgy, hogy ha $\eta \sim \mathcal{N}(0,1)$, akkor $\mathsf{P}(\eta > u_{\varepsilon/2}) = \varepsilon/2$. Ekkor

$$\mathsf{P}_{m_0}\left(\mathsf{u} \in [-u_{\varepsilon/2}, u_{\varepsilon/2}]\right) = 1 - \varepsilon,$$

azaz $1-\varepsilon$ valószínűséggel az u statisztika a $[-u_{\varepsilon/2}, u_{\varepsilon/2}]$ intervallumba esik.

Döntsünk a következő módon:

- Ha az u statisztika értéke a $[-u_{\varepsilon/2}, u_{\varepsilon/2}]$ intervallumba esik, akkor **elfogadjuk** a H_0 nullhipotézist, azaz a $[-u_{\varepsilon/2}, u_{\varepsilon/2}]$ intervallum az **elfogadási tartomány**.
- Ha az u statisztika értéke nem esik bele a $[-u_{\varepsilon/2}, u_{\varepsilon/2}]$ intervallumba, akkor **elvetjük** a H_0 nullhipotézist, azaz a $[-u_{\varepsilon/2}, u_{\varepsilon/2}]$ intervallum komplementere a **kritikus tartomány**.

4.1.1. Definíció. ***Elsőfajú hiba:*** *annak a valószínűsége, hogy a H_0 nullhipotézist elvetjük, pedig H_0 igaz. (Ezt szokás a próba terjedelmének is nevezni.)*

Másodfajú hiba: *annak a valószínűsége, hogy a H_0 nullhipotézist elfogadjuk, pedig H_0 nem igaz.*

Erőfüggvény: *annak a valószínűsége, hogy nem követünk el másodfajú hibát, azaz a H_0 nullhipotézist elfogadjuk, amikor az igaz.*

A próbát ***konzisztensnek*** *nevezzük, ha a másodfajú hiba 0–hoz konvergál (vagyis az erőfüggvény értéke 1–hez tart), amikor a minta elemszáma végtelenhez tart.*

A fenti döntési módszer esetén az elsőfajú hiba éppen ε, hiszen ha H_0 igaz, akkor

$$\mathsf{P}_{m_0}\left(\mathsf{u} \notin [-u_{\varepsilon/2}, u_{\varepsilon/2}]\right) = \varepsilon$$

Nyilván

$$\mathsf{u} = \frac{\overline{\xi}_n - m_0}{\sigma/\sqrt{n}} = \frac{\overline{\xi}_n - m}{\sigma/\sqrt{n}} + \frac{m - m_0}{\sigma/\sqrt{n}},$$

ahol

$$\frac{\overline{\xi}_n - m}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1),$$

ezért a másodfajú hiba valószínűsége

$$\begin{aligned}\mathsf{P}_m\left(\mathsf{u} \in [-u_{\varepsilon/2}, u_{\varepsilon/2}]\right) &= \mathsf{P}_m\left(\frac{\overline{\xi}_n - m}{\sigma/\sqrt{n}} \in \left[-u_{\varepsilon/2} - \frac{m - m_0}{\sigma/\sqrt{n}}, u_{\varepsilon/2} - \frac{m - m_0}{\sigma/\sqrt{n}}\right]\right) \\ &= \Phi\left(u_{\varepsilon/2} - \frac{m - m_0}{\sigma/\sqrt{n}}\right) - \Phi\left(-u_{\varepsilon/2} - \frac{m - m_0}{\sigma/\sqrt{n}}\right),\end{aligned}$$

ahol Φ a standard normális eloszlás eloszlásfüggvénye, azaz

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-v^2/2}\,\mathrm{d}v.$$

Ezért az erőfüggvény:

$$E_n(m) = 1 - \Phi\left(u_{\varepsilon/2} - \frac{m - m_0}{\sigma/\sqrt{n}}\right) + \Phi\left(-u_{\varepsilon/2} - \frac{m - m_0}{\sigma/\sqrt{n}}\right).$$

A próba konzisztens, azaz rögzített m esetén $\lim\limits_{n\to\infty} E_n(m) = 1$ (ezért $n \to \infty$ esetén a másodfajú hiba 0–hoz konvergál), ugyanis $\lim\limits_{x\to\infty} \Phi(x) = 1$, $\lim\limits_{x\to-\infty} \Phi(x) = 0$. Továbbá rögzített n mintaelemszám esetén $\lim\limits_{m\to\infty} E_n(m) = 1$ és $\lim\limits_{m\to-\infty} E_n(m) = 1$ (tehát $m \to \infty$ vagy $m \to -\infty$ esetén a másodfajú hiba 0–hoz konvergál). Az E_n függvénynek minimuma van az m_0 pontban, melynek értéke ε, ezért a másodfajú hiba nagy lesz (mégpedig $1 - \varepsilon$ közeli), ha m közel van m_0–hoz.

Nyilván ε csökkentése esetén az elsőfajú hiba valószínűsége csökken, viszont $u_{\varepsilon/2}$ növekszik, ezért a másodfajú hiba növekszik. (Tehát a kétféle hiba valószínűsége ellentétes irányban mozog!)

4.2. Egymintás t–próba

Tegyük fel, hogy $\xi_1, \xi_2, \ldots, \xi_n$ független, $\mathcal{N}(m, \sigma^2)$–eloszlású valószínűségi változók, ahol σ^2 és m is ismeretlenek. Legyen $m_0 \in \mathbb{R}$ egy rögzített szám. Legyen

$$\begin{cases} H_0: & m = m_0 \quad \text{nullhipotézis,} \\ H_1: & m \neq m_0 \quad \text{alternatíva.} \end{cases}$$

Tekintsük a

$$\mathsf{t} := \frac{\overline{\xi}_n - m_0}{\sqrt{s_n^{*2}/n}}$$

statisztikát. Ha a H_0 nullhipotézis igaz, akkor a 2.4.1. Tétel alapján a

$$\frac{\overline{\xi}_n - m_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1), \qquad \frac{n}{\sigma^2} s_n^2 = \frac{n-1}{\sigma^2} s_n^{*2} \sim \chi^2_{n-1}$$

statisztikák függetlenek, ezért

$$\mathsf{t} = \frac{\dfrac{\overline{\xi}_n - m_0}{\sigma/\sqrt{n}}}{\sqrt{\left(\dfrac{n-1}{\sigma^2} s_n^{*2}\right) \Big/ (n-1)}} \sim \mathsf{t}_{n-1},$$

azaz t_{n-1}–eloszlású. Legyen $\varepsilon \in (0,1)$ rögzített. Válasszuk meg a $t_{\varepsilon/2} \in (0, \infty)$ számot úgy, hogy $\mathsf{P}(\mathsf{t}_{n-1} > t_{\varepsilon/2}) = \varepsilon/2$. Ekkor

$$\mathsf{P}_{m_0}\big(\mathsf{t} \in [-t_{\varepsilon/2}, t_{\varepsilon/2}]\big) = 1 - \varepsilon,$$

azaz $1-\varepsilon$ valószínűséggel a t statisztika a $[-t_{\varepsilon/2}, t_{\varepsilon/2}]$ intervallumba esik.

Döntsünk a következő módon:

- Ha a t statisztika értéke a $[-t_{\varepsilon/2}, t_{\varepsilon/2}]$ intervallumba esik, akkor elfogadjuk a H_0 nullhipotézist, azaz a $[-t_{\varepsilon/2}, t_{\varepsilon/2}]$ intervallum az elfogadási tartomány.
- Ha az t statisztika értéke nem esik bele a $[-t_{\varepsilon/2}, t_{\varepsilon/2}]$ intervallumba, akkor elvetjük a H_0 nullhipotézist, azaz a $[-t_{\varepsilon/2}, t_{\varepsilon/2}]$ intervallum komplementere a kritikus tartomány.

Nyilván az elsőfajú hiba éppen ε, hiszen ha H_0 igaz, akkor

$$\mathsf{P}_{m_0}\big(\mathsf{t} \notin [-t_{\varepsilon/2}, t_{\varepsilon/2}]\big) = \varepsilon.$$

4.3. Kétmintás u–próba

Tegyük fel, hogy a $\xi_1, \xi_2, \ldots, \xi_k$ $\mathcal{N}(m_\xi, \sigma_\xi^2)$–eloszlású, és az $\eta_1, \eta_2, \ldots, \eta_\ell$ $\mathcal{N}(m_\eta, \sigma_\eta^2)$–eloszlású valószínűségi változók függetlenek, ahol σ_ξ^2 és σ_η^2 ismertek, m_ξ és m_η ismeretlenek. Legyen

$$\begin{cases} H_0: & m_\xi = m_\eta \quad \text{nullhipotézis,} \\ H_1: & m_\xi \neq m_\eta \quad \text{alternatíva.} \end{cases}$$

Ha a H_0 nullhipotézis igaz, akkor a 2.4.1. Tétel alapján

$$\mathsf{u} := \frac{\overline{\xi}_k - \overline{\eta}_\ell}{\sqrt{\dfrac{\sigma_\xi^2}{k} + \dfrac{\sigma_\eta^2}{\ell}}} \sim \mathcal{N}(0,1).$$

Legyen $\varepsilon \in (0,1)$ rögzített. Válasszuk meg az $u_{\varepsilon/2} \in (0,\infty)$ számot úgy, hogy ha $\eta \sim \mathcal{N}(0,1)$, akkor $\mathsf{P}(\eta > u_{\varepsilon/2}) = \varepsilon/2$. Ekkor megint vehetjük elfogadási tartománynak a $[-u_{\varepsilon/2}, u_{\varepsilon/2}]$ intervallumot, kritikus tartománynak pedig ennek a komplementerét.

4.4. Kétmintás t–próba

Tegyük fel, hogy a $\xi_1, \xi_2, \ldots, \xi_k$ $\mathcal{N}(m_\xi, \sigma_\xi^2)$–eloszlású, és az $\eta_1, \eta_2, \ldots, \eta_\ell$ $\mathcal{N}(m_\eta, \sigma_\eta^2)$–eloszlású valószínűségi változók függetlenek, ahol σ_ξ^2, σ_η^2, m_ξ és m_η is ismeretlenek. Legyen

$$\begin{cases} H_0: & m_\xi = m_\eta \quad \text{nullhipotézis,} \\ H_1: & m_\xi \neq m_\eta \quad \text{alternatíva.} \end{cases}$$

Ha a H_0 nullhipotézis igaz, akkor a 2.4.1. Tétel alapján

$$\mathsf{t} := \frac{\overline{\xi}_k - \overline{\eta}_\ell}{\sqrt{\left(\dfrac{1}{k} + \dfrac{1}{\ell}\right) \dfrac{(k-1)s_{\xi,k}^{*2} + (\ell-1)s_{\eta,\ell}^{*2}}{k+\ell-2}}} \sim \mathsf{t}_{k+\ell-2}.$$

Legyen $\varepsilon \in (0,1)$ rögzített. Válasszuk meg a $t_{\varepsilon/2} \in (0,\infty)$ számot úgy, hogy $\mathsf{P}(\mathsf{t}_{k+\ell-2} > t_{\varepsilon/2}) = \varepsilon/2$. Ekkor megint vehetjük elfogadási tartománynak a $[-t_{\varepsilon/2}, t_{\varepsilon/2}]$ intervallumot, kritikus tartománynak pedig ennek a komplementerét.

4.5. F–próba

Tegyük fel, hogy a $\xi_1, \xi_2, \ldots, \xi_k$ $\mathcal{N}(m_\xi, \sigma_\xi^2)$–eloszlású, és az $\eta_1, \eta_2, \ldots, \eta_\ell$ $\mathcal{N}(m_\eta, \sigma_\eta^2)$–eloszlású valószínűségi változók függetlenek, ahol σ_ξ^2, σ_η^2, m_ξ és m_η is ismeretlenek. Legyen

$$\begin{cases} H_0: & \sigma_\xi^2 = \sigma_\eta^2 \quad \text{nullhipotézis,} \\ H_1: & \sigma_\xi^2 \neq \sigma_\eta^2 \quad \text{alternatíva.} \end{cases}$$

Tekintsük az

$$\mathsf{F} := \frac{s_{\eta,\ell}^{*2}}{s_{\xi,k}^{*2}}$$

statisztikát. Ha a H_0 nullhipotézis igaz, akkor a 2.4.1. Tétel alapján

$$\mathsf{F} = \frac{\left(\dfrac{\ell-1}{\sigma_\eta^2} s_{\eta,\ell}^{*2}\right) \Big/ (\ell-1)}{\left(\dfrac{k-1}{\sigma_\xi^2} s_{\xi,k}^{*2}\right) \Big/ (k-1)} \sim \mathsf{F}_{\ell-1,k-1},$$

azaz $\mathsf{F}_{\ell-1,k-1}$–eloszlású. Legyen $\varepsilon \in (0,1)$ rögzített. Válasszunk olyan $0 < c_1 < c_2$ számokat, hogy $\mathsf{P}(\mathsf{F}_{\ell-1,k-1} < c_1) = \mathsf{P}(\mathsf{F}_{\ell-1,k-1} > c_2) = \varepsilon/2$. Ekkor vehetjük elfogadási tartománynak a $[c_1, c_2]$ intervallumot, kritikus tartománynak pedig ennek a komplementerét.